10/030062

4/PRTS

531 Rec'd PCT/PTO 04 JAN 2002

Description of the Invention

METHOD FOR THE MULTI-DIMENSIONAL ANALYSIS OF A PROTEOME

The invention is directed to a method for the multidimensional analysis of a proteome in which the biological tissue with the proteome to be analyzed is solubilized and the proteins belonging to the proteome are separated, quantitatively determined and identified. The method is used in biochemistry, biotechnology, medicine and in the pharmaceutical industry for purposes including diagnostics and the development of biologically active substances. Special areas of use are opening up in fundamental research, e.g., for clarifying questions pertaining to developmental biology or cell differentiation and in related research for screening active ingredient banks, for the development and optimization of biologically active substances or for differentiating between normal and pathogenic states in organisms.

Recently, genomes of organisms have been sequenced completely or in large part [Fraser, C. M. et al.: The minimal gene complement of Mycoplasma genitalium, Science, 1995, Oct. 20, 270 (5235), 397-403; Fleischmann, R. D. et al.: Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 1995, July 28, 269 (5223), 496-512; Blattner, F. R. et al.: The complete genome sequence of Escherichia coli K-12, Science, 1997, Sept. 5, 277 (5331), 1453-74; Goffeau, A. et al.: Life with 6000 genes, Science, 1996, Oct. 25, 274 (5287), 546, 563-7]. Sequencing of cDNA portions has been even more intensive [Clark, M. S.: Comparative genomics: the key to understanding the Human Genome Project, Bioessays, 1999, Feb. 21 (2), 121-30; Evans, M. J. et al.: Gene trapping and functional genomics, Trends Genet., 1997, Sept. 13 (9), 370-4]. The sequence data are stored in databases. The clarification of the genome of an organism ultimately leads "only" to an understanding of the relatively static information content of the genetic material for this organism. With cDNA sequences, it is possible, in principle, to determine expression levels of the mRNA as they relate to specific cells and specific environments and accordingly to obtain a gene expression pattern of the RNA.

From a gene of the genome, it is possible a) to develop by different processes various mRNA types which code for divergent proteins, and b) to form a large number of extremely differently functioning proteins from this by means of posttranslational modification. Previously known modifications include phosphorylation and dephosphorylation, limited proteolysis, acetylation, methylation, adenylation, sulfation, glycosylation [McDonald, L. J., et al.: Enzymatic and nonenzymatic ADP-ribosylation of cysteins, Mol. Cell. Biochem., 1994 Sept., 138 (1-2), 221-6; Baenziger, J. U.: Protein-specific glycosyltransferases: how and why they do it!, FASEB J., 1994, Oct. 8 (13), 1019-25; Mimnaugh, E. G. et al.: The measurement of ubiquitin and ubiquitinated proteins, Electrophoresis, Feb. 1999, 20 (2), 418-28; Davis, P. J. et al.: Protein modification by thermal processing, Allergy, 1998, 53 (46 Suppl.), 102-5]. However, the expressed and modified proteins ultimately yield the pattern which describes the cell differentiation and the reaction to internal and external influences of cells. Most striking is the limited importance of knowing the genome for the realization of a defined biological state when the various cells in different organs and inside an organ are compared. For example, a liver paranchyma cell, a nerve cell of the brain and a mucosa cell of the intestine have the same set of genetic information but completely different functions brought about by the regulation of the expression of the genome in these cells and the regulation of the enzyme pattern and protein pattern within the cells and the various tissues.

DNA	RNA	Proteins
Static and descriptive, with exceptions	Transfer of information. Quantity is regulated and transfers the information of the DNA to the protein plane.	Maintaining cell structure, reaction to changes and signals. Interactions with other cells. Quantity and activity are regulated.

The term "proteome" was first used in 1996 [Friedrich, G. A.: Moving beyond the genome projects, Nat. Biotechnol., Oct. 1996, 14 (10), 1234-7].

The proteome, that is, the totality of all proteins in a cell, with a definite development stage and under defined environmental conditions, is a much more dynamic representation of the physiological state of cells, organs and organisms. Proteome analysis investigates which parts of the genome are expressed and modified under defined, cell-specific conditions. This has led to rapidly growing interest in this field, leading to a growing number of publications (PubMed search term: Proteome; over the last 1 year: 64 hits; over the last 2 years: 99 hits; over the last 5 years: 122 hits), conferences and events on this subject.

In order to obtain a quantifiable "picture" of a proteome, the following procedure is currently performed: In a first step, the biological materials must be solubilized and homogenized (exceptions: e.g., in a serum, they are in a homogenous solution). The proteins are isolated or separated in the second step and identified in the third step. In the fourth step, the obtained data are evaluated [Ben, R. H., et al.: Two dimensional electrophoresis, The state of the art and future directions, Proteome Research, New frontiers in functional genomics, Springer 1997, Chap. 2, 13-33].

1. Solubilization

Methods and arrangements known from biochemistry are used for this purpose, e.g., shear homogenizers, ultrasonic processing, high-pressure pressing. The difficulty consists in quantitative solubilization which does not destroy the function of the proteins as far as possible, because only quantitatively solubilized proteins provide a real picture of the specimen material in the subsequent second step (separation and detection of proteins) [Rabilloud, T.: Solubilization of proteins in 2-D electrophoresis, An outline, Methods Mol. Biol;., 1999, 112, 9-19; Rabilloud, T. et al.: Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients, Electrophoresis, Mar.-Apr. 1997, 18 (3-4), 307-16; Staudenmann, W. et al.: Sample Handling for proteome analysis, Electrophoresis, May 1998, 19 (6), 901-8].

2. Separation and detection

At present, two-dimensional gel electrophoresis is essentially used for separating the proteins of the proteome. First tests with two-dimensional HPLC have been carried out. However, they have not yet achieved the separation effect of two-dimensional electrophoresis [Opiteck G. J. et al.: Comprehensive two-dimensional high-performance liquid chromatography for the separation of overexpressed proteins and proteome mapping, Anal. Biochem. May 1998, 1; 258 (2): 349-61]. The first dimension of two-dimensional electrophoresis is isolation according to the isoelectric point, that is, ultimately, according to the charge characteristics of a protein. In the second dimension, the proteins are separated according to size in a denaturing sodium dodecyl sulfate gel. This separation technique has been known for about 20 years. An advantage of two-dimensional electrophoresis consists in the possibility of separating a relatively large number of proteins on a surface with high resolution. Presently, it is assumed that approximately 10,000 proteins can be detected in a two-dimensional gel of this kind [Klose, J. et al.: Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome, Electrophoresis, 1995, June 16 (6), 1034-59]. Another advantage is that it is possible to quantify the separated proteins by radioactive marking or after staining with techniques that are likewise known. These quantification methods are protein-specific, have a limited dynamic detection range, are generally difficult to automate and are dependent on the respective conditions of use (which often can not be reproduced) [James, P.: Of genomes and proteomes, Biochem. Biophys. Res. Commun., 1997, Feb. 3, 231 (1), 1-6]. They are only suitable for relative determinations. Quantification by immunological characteristics is problematic because blot techniques having limited meaningfulness in terms of quantitative information must be used for this purpose.

This results in a fingerprint-like pattern which characterizes the proteome.

This separation technique has the following disadvantages:

- limited dynamic range due to the load capacity of the separating gel
- the maximum quantity of proteins that may be used is limited to a range of µg to mg protein [James, P.: Of genomes and proteomes, Biochem. Biophys. Res. Commun., 1997, Feb. 3, 231 (1), 1-6]
- restriction of sample volume used
- separation is limited to two dimensions

- the ampholytes required for separation and the acrylamide gel material can lead to artifacts and can accordingly contribute to misinterpretations which are difficult to detect
- proteins that are present in very high concentrations result in relatively strong signals and overlap proteins in low concentrations, so that direct identification and quantification is impossible in this case
- the loss of the native conformation in denaturing separating gel causes the loss of biologically functional characteristics and impedes the identification of proteins by determining their biological characteristics, for example, their catalytic activity or specific bonding characteristics
- secondary analysis, such as the frequently used specific proteolysis of individual proteins, followed by determinations of mass necessitates a step for extracting from the gel or blot membrane which is difficult to automate.

3. Identification of proteins

Sequencing, mass analysis and estimation of the isoelectric point from the length of run in the gel and mass analysis of peptide fragments after separation from the gel and tryptic digestion in mass spectrometry are normally used for this purpose [Shevchenko, A. et al.: Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two-dimensional gels, Proc. Natl. Acad. Sci. USA, 1996, Dec. 10, 93 (25), 14440-5; Traini, M. et al.: Towards an automated approach for protein identification in proteome projects, Electrophoresis, 1998, Aug. 19 (11), 1941-9]. Features such as the catalytic activity of the proteins and the native conformation are almost completely excluded from the utilized separating technique and are not available for identification.

In particular, the known identification methods have the following advantages and disadvantages:

- The sequencing is carried out by Edman degradation in automated arrangements and is relatively costly and time-consuming. It requires greater quantities of the protein. Therefore, in spite of current further development for mass screening, it is less suitable [Gooley, A. A. et al.: A role for Edman degradation in proteome studies, Electrophoresis, 1997, June 18(7), 1068-72]. However, this analytic step is necessary in most cases for identification of primarily unknown proteins.

- The specificity of information of mass determination of a protein which should finally lead to its identification is increased in that the proteins undergo protease digestion after separation and the information obtained by means of mass analysis is compared with the masses of the peptide sequences predicted from the primary structure after tryptic digestion. Essentially two types of mass spectrometry are used: The first is Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI-MS) and the second is ElectroSpray Ionization Mass Spectrometry (ESI-MS) [Ducret, A. et al.: High Throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry, Protein Sci., 1998, Mar 7 (3), 706-19; Parker, K. C. et al.: Identification of yeast proteins from two-dimensional gels: working out spot cross-contamination, Electrophoresis, 1998, Aug. 19 (11), 1920-32]. The first method has the advantage that it allows a very large mass range of up to 1 million Dalton to be analyzed and can be carried out in a relatively robust manner. However, it can be carried out only discontinuously. The ESI technique, on the other hand, can be appended almost continuously to separating techniques and is presently showing a sharp growth in the development of breadth of application as well as technological possibilities. The enormous advances achieved in recent years with both techniques allow mass resolutions to isotope distribution, that is, resolutions of less than 1 Dalton. In this way, a mass spectrum of peptide fragments is obtained according to sequence-specific, defined protease digestion or another defined splitting of the proteins. This spectrum is typical for every protein and is used for protein identification in sequence databases of proteins and expressed sequence tag banks. Since the identification of the protein by precise identification of the predicted peptides takes place after protease digestion, any posttranslational modification of the proteins, e.g., by glycosylation, interferes with detection. Further, fragmentation spectra of the individual peptides in the mass spectrometer can supply information about the amino acid sequence of the peptides. This sequence information can be used by itself or along with the other known protein data to identify this protein in a sequence database. This method of sequence analysis is not yet routinely used at present due to the difficulties of correct data interpretation. The limits of protein identification through mass spectrometry methods reside in the incomplete detection of all protein sequences in existing databases.

4. Data analysis

The characteristics of the individual detected proteins from separation in two-dimensional electrophoresis, such as quantity, isoelectric point and size, and the data for protein identification from additional steps, e.g., sequencing or mass spectrometry, are combined. This produces the picture of the totality of the proteins with their identity and quantity in the respective proteome.

It is the object of the invention to improve and facilitate quantification and identification of the proteins of a proteome and to make it possible for certain proteins to be quantified and identified for the first time.

According to the invention, the proteins of the proteome are subjected to a number n of different separating processes under standardized conditions in such a way that each of the m_1 liquid fractions obtained in a separating step supplies m_2 liquid fractions in a subsequent separating step, wherein, after n separating steps, there are $m_1 * m_2 * \ldots m_n = M$ liquid fractions which are identified by τ different analysis processes qualitatively and/or quantitatively by identification processes, known per se, and determined quantitatively by quantification processes which are likewise known per se, so that after combining the analysis data an n-dimensional image of the proteome is obtained which is characterized by identifiers and quantifiers and by the position in the n-dimensional data space.

Advantageous embodiment forms of the method are set forth in the subclaims 2 to 12.

The method according to the invention is not subject to the tight limitation on quantity due to the load capacity of previously used two-dimensional electrophoresis. Protein quantities in the range of several grams can be used. The separating matrices can be utilized repeatedly. In this way, greater reproducibility of results can be achieved. The sample material that is used is in liquid phase and is accordingly immediately accessible for subsequent analysis steps. The improved maintaining of native characteristics during separation makes possible analytic methods such as activity determination and immunological processes based on the native conformation of the analytes. The separation of analytes with the same charge characteristics and size characteristics is not possible in the two-dimensional electrophoresis that is usually used. However, this restriction is eliminated through the use of at least one further characteristic, such as the hydrophobicity of the analytes, for separation. After separation, the samples in fractions are also available for additional preparative tasks.

The invention will be described more fully in the following with reference to an embodiment example shown in the drawing.

Fig. 1 shows the separation of 1000 proteins in three dimensions

Fig. 1a:	fractions 1 to 33
Fig. 2a:	fractions 33/34 to 67
Fig. 3a:	fractions 68 to 100;

Fig. 2 shows a graphic three-dimensional view of the fractions according to Fig. 1.

As an embodiment example, 1000 proteins are to be described by three characteristics A, B, C. These characteristics may be, e.g., size, charge and hydrophobicity. The characteristics are randomly distributed in the proteins. All proteins are numbered consecutively. Subsequently, separation is carried out according to characteristic A (e.g., size), resulting in 100 fractions a with the corresponding proteins. These fractions a are separated into 10 fractions b according to characteristic B (e.g., charge).

Each of these fractions b is subjected to separation based on characteristic C (e.g., hydrophobicity) and gives fractions c. In total, 100 x 10 x 10 = 10,000 individual fractions are obtained. Each protein obtained by separation is uniquely allocated to one of the fractions a, b, c according to its characteristics. In the assignment according to Fig. 1, the respective fractions are designated by number. In this case, the fractions a are associated with characteristic A. They divide the possible value range of characteristic A into one hundred equal parts, i.e., assuming a value range from 0 to 100, value 1, for example, corresponds to range 0 to 1, value 2 corresponds to range 1 to 2, …, and value 100 corresponds to range 99 to 100. Analogously, the possible value ranges of characteristics B and C are divided into ten equal parts, i.e., value 1, for example, corresponds to range 1 to 10. On the average, every tenth fraction contains a protein.

Considered at random, there is a possibility of multiple assignments. In the example shown in the list according to Figs. 1a-c, there are 39 double occupancies and one triple occupancy of fractions.

For reasons of space and for the sake of clarity, the empty 9,000 fractions are not shown.

Fig. 1 contains the following list in tabular form:

Protein No.	Fractions a	Fractions b	Fractions c

Fig. 1a shows fractions a = 1 to 33, Fig. 2a shows fractions a = 33/34 to 67 and Fig. 1c shows fractions a = 68 to 100. Fig. 2 shows a three-dimensional diagram with the positions of the fractions occupied by proteins according to Fig. 1.

Assignment of Reference Numbers

A, B, C - characteristic of proteins

a, b, c - fraction